DOCUMENT RESUME

ED 147 757 CS 003 631

AUTHOR Mione, Stephen A.
TITLE Criterion-Referenced Testing: A Critical Perspective.
PUB DATE 77
NOTE 27p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Basic Skills; Behavioral Objectives; *Criterion Referenced Tests; Elementary Secondary Education; Individualized Instruction; *Literature Reviews; Measurement Techniques; *Performance Based Education; *Test Construction; *Testing Problems

ABSTRACT

This paper reviews the literature related to criterion-referenced testing (CRT), which has been tied to trends in individualized instruction, competency-based education, and the accountability movement. Among the conclusions that are drawn are (1) CRT scores assume a good match between test items and behavioral objectives that the CRT attempts to measure, and, to the extent that the test items do not accurately measure the objectives, the results will be inaccurate; (2) the theory on determining mastery levels is rather fragmentary and inconclusive; (3) reliability and validity studies on CRT are much needed; (4) CRT measurement should be based on a threshold loss function rather than a squared-error loss function; (5) Bayesian statistical procedures should be used to improve proportion-correct estimates; and (6) customized criterion-referenced tests need to be developed by local school districts by contracting with testing organizations. It is concluded that there is a need for awareness of both the strengths and weaknesses of criterion-referenced testing. (AA)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED147757

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

CRITERION-REFERENCED TESTING:

A CRITICAL PERSPECTIVE

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Stephen A. Mione

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

Stephen A. Mione

CS 003 631

Criterion-Referenced testing is tied to several trends: individualized instruction, competency-based education, and the accountability movement. Governor Brendan Byrne signed into New Jersey state law statewide minimum educational standards for all public schools. "Project 81" in Pennsylvania is a move towards a competency based educational program. A recent article in the Philadelphia Inquirer (11/27/76) states: "Indeed, the latest Gallup Poll on education showed that two-thirds of the American public, 65 percent, think high school seniors should be required to pass a standardized examination before being eligible for a diploma" (p. 1A). The accountability movement "Back to Basics" proponents, decline of national test scores, rising interest in applied performance testing along with a concern about the quality of a high school diploma have led, since January of 1976, to thirty-three states giving consideration to the establishment of some form of minimal competency testing.* Competency or performance tests are usually criterion-referenced tests: a test that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards. Criterion-referenced tests are not designed to determine an individual's relative standing in some norm group as do most standardized achievement tests. Instead,

*As of January 1975, Representative Albert H. Quie (R-Minn.) proposed an Amendment to title I of the 1965 Elementary and Secondary Education Act which is pending before Congress. The Amendment would establish criterion-referenced testing as a means of determining how much money a state would receive from the federal government. The distribution would be made according to each state's share of the total number of educationally disadvantaged children identified by the tests.

criterion-referenced tests indicate what an individual can or cannot do with certain specific requirements specified as a criterion score which does not depend on the scores the other students actually obtain on the test. There is, consequently, no limitation on how many students may receive each grade (A through F) and a student may retake a criterion-referenced test until a mastery score is achieved.

The values of criterion-referenced testing are heralded by some educators as the key to individualization. Adopting the very sound premise that students learn at different rates CRT (criterion-referenced testing) fits well with programs emphasizing individualized instruction. Some educators further argue that competition can be eliminated and student learning can be greatly facilitated, because CRT forces one to state one's teaching aims in the form of behavioral objectives. Behavioral objectives become the absolute standards to which test performance on a CRT is related.

A review of the research literature on criterion-referenced testing, however, reveals that there are major problems related to CRT. One major problem area is the issue of "how to determine mastery?" which leads into two sub-topics--how to determine true scores and how to establish cut-off scores. A second problem deals with the availability and construction of criterion-referenced tests. Finally, a third major issue is that of the weaknesses inherent in a behavioral objectives approach to teaching and testing. This paper seeks to review the research, in general, on criterion-referenced

testing and to focus on the aforementioned problem areas not only as the research clarifies the problems but also as this writer evaluates the literature and adds his own critical perspective. The thesis of this paper, after a review of the research on criterion-referenced testing, is that a CRT is not inherently good or bad just as a norm-referenced test is not inherently good or bad. Criterion-referenced tests are intended to complement norm-referenced measures in the assessment of a pupil's skills. The decision to use a norm-referenced test or a criterion-referenced test should be based upon the purpose of testing. The shortcomings and advantages of both measures should be foremost in the development of tests and the interpretation of the results. As this reader views criterion-referenced tests, they are not the panacea that some educators suggest in the areas of competency-based education, individualized instruction, and educational accountability.

When reviewing the literature on CRT there is a thread in the research dealing with grading and competition. Advocates of CRT argue that norm-referenced tests encourage cut-throat competition, negativeness, dishonesty, and grade-grubbing (Power, 1976). The lack of sound educational learning theory is also posed as a problem underlying the use of norm-referenced testing. Students learn in different ways and at different rates, therefore, it is meaningless to teach to the average student, usually combining only a couple of

basic teaching methods (lecture and question-answer)(Aubertine, 1972) and then to compare a student's achievement to the norm (Hambleton, 1974; Gronlund, 1974). The CRT proponents tell us to fit the test to the instruction (Wittrock, 1970), to determine whether the individual has attained a prescribed degree of competence on the performance task. The question of precise achievement levels and comparisons among individuals on these levels are irrelevant. An individualization process is suggested (Hambleton, 1974) whereby teachers and students work together to make decisions about the selection of specific instructional objectives, to develop instructional prescriptions, and to select instructional resources. When the student feels he has mastered the objectives, a CRT is administered. If mastery is not achieved, recycling occurs: re-diagnosing, re-teaching, re-learning, and then re-testing until mastery is achieved. Thus, a student learns at his own rate and, hopefully, learning materials and methods will be selected to suit the pupil's learning style.

As related to the premise that students learn at different rates the "time variable" in learning tasks becomes an issue (Carroll, 1970). Carroll introduces a time continuum:

Learning Task start _____ Criterion Attainment
time

and suggests that some students may never achieve the criterion. Ability, aptitude, motivation, interest, perseverance, oppor-

tunity to learn (teacher time allowed), quality of instruction are variables that may prevent student mastery of objectives no matter how much time is spent on teaching and studying. Carroll also suggests that measurement of learning is a linear function of the ratio $\frac{\text{time taken to learn}}{\text{time needed to learn}}$ is weak. The assumption is that it is possible to measure amounts of gain in equal units--a very shaky assumption. Also, a complicating factor is a student's learning style and aptitude inter acting with style of instruction. A priori learnings also make such a linear function theory questionable. Carroll's perspective is important to consider as it relates to the time factor and suiting instruction to individual student's rate of learning.

Washburne (1922) also introduced the time factor in learning as early as 1922: "...time should be the variable unit, whereas achievement should be the constant one." Packard (1972) states that measurement strategy should incorporate differential weighting to equalize successive curricular sub-units, e.g. time/effort indices to pages of units. The strategy suggested is that the Basic Rate of Measure is of the form: R/Min. (where "R," the response, is a small assignment of one of several sub-objectives) and (where "Min." is the time taken by the student to complete the assignment to mastery of criterion). Thus, each data point equals R/Min. = 1/Min. The basic rate is weighed by making it the numerator of a fraction where the denominator is a rate of the form:

R/S where "R" is the same unit(s) assigned to the student and "s" is a statistic, e.g. (mode) of the time taken by many students to compute the same unit(s). Since the "R" in both the numerator and denominator is one for each data point, this compound fraction resolves to the form: S/Min. There seems to be no apparent resolution of the issue of the time factor in learning in the literature on CRT.

Ebel (1971) provides a good point of transition to lead into the problem areas or "limitations" of CRT. Ebel writes a rather general article but he touches on the major weaknesses with CRT: the weaknesses of behavioral objectives, the problem of defining mastery scores, and the difficulty of CRT construction. Ebel states: "...the difficulties and limitations of criterion-referenced measures, which half a century ago led to their virtual abandonment, would once again become apparent and would, in all probability, start the pendulum swinging back toward norm-referenced measures"(p. 283)--that is, assumedly, if one would take seriously Ebel's points. Block writes a rejoinder that in a very specific fashion takes up each of Ebel's points against CRT and offers an alternative positive opinion. Neither writer is particularly convincing since their arguments are rather general.

The problem of determining mastery levels on a CRT is a complex issue and has received much attention in the research literature. Glaser (1963) proposes that a criterion-referenced measure is related to a student's acquisition of knowledge along a continuum from no proficiency to perfect performance.

Glaser suggests that criterion levels can be established at any point along the continuum and indicates what an individual can or cannot do. Minimum competencies should be established for the end of a course in order to progress to the next level course in a sequence. Later, Glaser and Nitko,(1971) clarified Glaser's 1963 position by stating that a CRT is one constructed to yield measurements which are directly interpretable in terms of specified performance standards.

Emrick (1971) presents a model based upon four basic assumptions, as follows, for determining mastery levels of a CRT:

1. Learning of fundamental skills can be considered all or none.

2. Each item response on a single skill test represents an unbiased sample of the examinee's true mastery status.

3. Measurement error occuring on the test can be of only one type for each examinee.

4. Through practical and theoretical considerations of evaluation error costs and item error characteristics, an optimal mastery criterion can be calculated.

Emrick's Model is clearly formulated:

1. Behavioral objectives are established and mastery of them will be a binary (all or none) variable. The level or degree of mastery of the objective is determined by the proportion or number of component skills which are mastered.

2. Test items are to be kept homogeneous in content, form, and difficulty level.

3. Measurement error of each skill is based on the binary system. There are two types:

a. Type I or alpha (α) in which the examinee's response leads to a mastery conclusion when true status is non-mastery, e.g. lucky guesses, or

b. Type II or beta (β) in which the item response leads to a non-mastery conclusion when, in fact, the examinee has mastered the skill, e.g. careless errors.

Consequently, mastery or non-mastery exists.

4. The extent of measurement error in a single skill test can be approximated by calculating the average inter item correlation of examinee responses to the parallel and homogeneous test items. This average inter item correlation provides an unbiased estimate of the squared correlation between an examinee's true mastery state and his item response.

5. Due to measurement error, decision errors will accrue regarding determination of examinee status on skills. Emrick advocates a decision-theoretic approach to this problem which in turn suggests a cost-benefit analysis of the variables which comprise the evaluative process. Three classes of these variables are the following:

a. Statistical (item reliability and test length)

b. Curricular (objectives viewed as ancillary will be low in regret can become sizeable for objectives fundamental to the model)

(1) Type I--false pass

(2) Type II--false fail

(3) Based upon above the cost in re-teaching and retesting

c. Psychological Costs

(1) Type I--Confusion, sub-optimal success rate

(2) Type II--boredom, decreased sense of achievement, low motivation

6. Mastery Cut-Off Scores can be optimized in terms of relative decision error costs and relative item error probabilities according to the following formula:

$$\log \frac{\beta}{1-\alpha} + 1/n \ (\log RR)$$

$$K = \log \frac{}{(1-\alpha)(1-\beta)}$$

where

K = the cut point expressed as a percent score on the test

α = estimated probability of Type I item error

β = estimated probability of Type II item error

RR = Ratio of regret of Type II to Type I decision errors

N = test length (number of items)

Emrick's model seems to ignore the in-between state of mastor and non-master whereby a student may have achieved some skills that allow for solution of some but not all of the problems in a task.

Meskauskas (1976) points out that during the 1960's advocacy for adopting CRT from an educational or philosophical

point-of-view proliferated in spite of reservations of classical measurement theorists. In the 1970's, Meskauskas states, measurement specialists are presenting papers of a mathematical nature with mathematically-based techniques for deciding upon cutting scores, mastery models and procedures for establishing the "cutting" or pass-fail point. The two issues: the definition or mastery and the issue of a priori standards are viewed as crucial by Meskauskas. Measurement specialists have presented two approaches: Decision theory and Bayesian Statistics. Emrick's decision theory model is viewed as "...seeking perfection in an imperfect world with imperfect measuring instruments" (p. 143). The consequence is that the decision-makers dealing with a CRT choose a level of mastery that "seems right" which is basically intuitive and arbitrary. Emrick's three variables that comprise the evaluative process: test factors (statistical), curricular and psychological are difficult to quantify empirically and is time-consuming when possible. Meskauskas goes on to conclude:

> Decision models that are not specifically designed for application to either a continuum or state mastery model are incomplete. There is a danger of misapplication of incompletely stated models, because a decision prodedure may be applied to an essentially incompatible mastery model, with a resulting erroneous decision. Also, the acquisition function should be useable as part of the prior knowledge available to the decision-maker. If this were quantifiable, it could help to improve accuracy of classification. (p. 156)

Livingston (1972) provides another point-of-view on

mastery levels in terms of error of measurement and true scores. A reliability coefficient for criterion-referenced tests is developed from the assumptions of classical test theory. This coefficient is based on deviations of scores from the criterion score, rather than from the mean. The coefficient is shown to have several of the important properties of the conventional norm-referenced reliability coefficient, including its interpretation as a ratio of variances and as a correlation between parallel forms, its relationship to test length, its estimation from a single form of a test, and its use in correcting for attenuation due to measurement error.

Harris (1972) provides an alternative interpretation to Livingston's reliability coefficient for a CRT. A notion of the relation of the size of the reliability coefficient to the range of talent is offered. It is shown that the larger Livingston coefficient does not imply smaller standard error of measurement and, consequently, does not imply a more dependable determination of whether or not a true score falls below or exceeds a given criterion value.

Millman (1973) deals with the determination of a passing score on a CRT that offers an alternative to the classical test theory, binomial, and sequential models which Millman views as "too technical". His suggestion is to set a passing score so that a predetermined percent of students pass. Test items should be analyzed and number of items and importance of each item established according to the following schema:

Difficulty or Importance: Level I -- all items correct

Level II -- some fraction correct

Level III -- smaller fraction correct

A Blue Print or Grid should be developed and minimal competencies and passing scores outlined following a "Probabilistic-Approach." Minimum cut off scores are viewed as important by Millman in sequential learning situations where too low a level for passing could have a negative effect on future learnings. He also considers the psychological and financial costs involved when one sets a low passing score but little resources exist for remedial educational programs.

Davis and Diamond (1974) present a Bayesian method of approaching mastery levels whereby a table estimating probability that a student's true competency level is at or above a number of chosen levels based upon an observed score. Successful classification rates are presented rather than error rates. For a five-item test, the probability is that a student with an obtained score of 5 has a true competency level of .99. Cutting scores at two levels of correct classification: .50 and .85 are made. The cutting score for tests up to 20 items, if there are 85% correct classifications of students with competence levels of 90% and higher, would be set equal to the number of items in the test. This goal would be attained until a test length of more than twelve but less than twenty items was used.

No valid and reliable data or conclusions have been reached on the issue of mastery levels, cut-off scores, error

of measurement indices of test effectiveness. It is clear that at present much more developmental work is necessary on this problem area by educational measurement specialists (Dziubian and Vickery, 1973).

A second problem, not as fully developed in the research literature as determining mastery levels, but probably equally important, is the issue of the difficulty of constructing criterion-referenced tests and their availability as suited to the needs of particular school districts, schools, and subject area teachers. The question becomes "Where do suitable criterion-referenced tests come from?" Will commercial test publishers produce high-quality and appropriate criterion-referenced tests? The answer seems to be--"NO!" Popham (1977) points to two possible reasons for the negative response: a CRT "...must unequivocally describe the domain(s) of examinee behavior it assesses" and "Except, perhaps, in reading and mathematics, the curricular emphases in U.S. school systems are so diverse as to render only a small proportion of their curricula appropriately assessable by a single criterion-referenced test" (p. 258). Consequently, it is almost impossible for a test publisher to develop a CRT that will coincide with local educational emphases.

What then is the answer? There are two logical and feasible solutions. Local districts can develop their own criterion-referenced tests. However, the skills necessary to develop a quality, reliable, valid CRT is no easy task. A specialist for a district would have to be hired or an

interested, capable employee of a local school district would have to receive specialized training. Popham suggests a second alternative--"Contractual Test Development." Testing organizations would contract with schools to create "customized criterion-referenced measures suitable for a particular school system"(p. 259). Popham has apparently given serious thought to such a concept. He is very much aware of the cost factor and suggests several ways to keep the costs as low as possible:

1. Get state departments of education to underwrite the cost of criterion-referenced tests in the basic skills areas of reading, writing and math--the widely desired competencies of the public--and then local districts could share the tests.

2. Limit the traits tested by a CRT to a modest number of highly important competencies.

It will be interesting to see what approaches the various states adopt as they move to a competency-based program in their attempts to measure minimal competencies.*

The final area of concern which suggests a problem with criterion-referenced testing is the total reliance a CRT must reveal in its relationship to the competencies it is testing mastery of as stated in the form of behavioral objectives. It is essential to the planning and improvement of instruction, to the evaluation of learning outcomes and, quite possibly,

*Oregon is the only state to have had a full-fledged competency based educational program for any length of time and their model should be worthwhile studying.

to individualizing instructing and to evaluating teacher effectiveness to use behavioral objectives. However, the weaknesses of the approach must also be recognized and a critical perspective must be realized on a conscious level when planning on educational program based upon behavioral objectives.

The assumptions upon which behavioral objectives are based are somewhat as follows. The student must be educated. Those responsible for educating should set objectives. To avoid ambiguity and vagueness in establishing the objectives, they should be specified in terms of the intended testable "outcomes" of teaching. Ideally, a properly defined behavioral objective should anticipate the question "How shall I know if the student has achieved the objectives of my course?" Behavioral objectives state the intended performances of the student.

The appeal is strong to the empirically oriented researcher on teaching, classroom learning and/or curriculum development since such objectives are amenable to observation and testing. There is an aura of scientific rigor and professionalism in the evaluation of programs and personnel which are based on a system akin to "scientific research." The appeal of scientific based education in its extreme is a view of human behavior that is reductionistic, deterministic, and physicalist. There are certain strengths and weaknesses built into these assumptions. I shall deal with the strengths first.

The strengths of the attempt to add a more scientific approach to evaluating learning outcomes are numerous. First, the teacher and students have clear purposes to guide instruction. Second, broad areas of content are broken down into manageable, meaningful units or pieces for teaching. Instead of a broad based objective such as "To appreciate literature," a more narrowed aim is stated -- "Given several appealing alternatives the student will voluntarily choose to read a piece of literature." Third, these more specific goals stated in behavioral terms can be organized into sequences and hierarchies that further facilitate learning. Fourth, evaluation of teacher effectiveness as reflected by student achievement of objectives is made an easier and simpler task. Finally, courses taught by different teachers to different students may share the same objectives. Thus, the comparative study of student success in learning and teacher success in teaching can be studied.

The presmises regarding behavioral objectives are compatible with the behavioral learning theory of the past fifty years. Furthermore, there is a strong current in educational research presently that aims at elevating instruction and instructional research to a level of scientific rigor similar to other behavioral and physical sciences. Behavioral objectives also certainly complement the technological advancements in education such as teaching machines and computer aided instruction. All of these factors place behavioral objectives in a most favorable light. The picture

will now be somewhat tinted by what I view as some of the chief weaknesses in a behavioral objectives approach to teaching-learning.

The chief weakness I see in the system is the emphasis placed on learning outcome. Isn't education more than merely mastering objectives? What of the process of education? It seems to me that a teacher ought to be equally concerned with "how the student learns" as with "what he learns." One of the most powerful elements in the process of schooling is social interaction, not content.

Also, proponents of behavioral objectives tend to value those aspects of content that can most easily be put into behavioral terms. The result is that basic skills can become over-emphasized and high level skills such as generalizing, abstracting, conceptualizing, problem solving, creative, critical thinking may be neglected because it is not always easily observable behavior. For example, how does the teacher state an outcome of observable student behavior for the following objectives:

To become a more sensitive person

To become a more humane individual

To develop a healthy, positive self-concept

To develop a sense of social responsibility

These are objectives dealing with a character trait or attitude. In these cases it is not clear just what a teacher is expected to teach and what a student is expected to learn.

Another factor is that even if some behavioral objectives could be stated for these affective goals, such development may not show observable growth in a year. They are more long-term objectives. They are also internalized rather than externalized goals. Pupils organize content psychologically as well as logically. Suppose a pupil has gained attitudinal growth but fails to reveal it in the behavioral terms narrowly selected by the teacher. Couldn't growth be revealed outside of class or in situations different from those established by the tester?

Another major weakness can be argued in terms of the stress behavioral objectives place upon the small units and hierarchy of "macro" and "micro" objectives in building a step-by-step, sequential pattern to teaching and learning. There is a problem of flexible adaptability. There is a danger that the behavioral objectives may compose a rigid formality in teaching. The teacher may feel locked into stated objectives when a new, equally important or more important objective may come to mind. The teacher may also feel little need for creative thought about achieving the objectives once formulated. Dynamic teaching may be lessened. Teaching for a CRT might become a sole aim for the instructor and students then could view mastery as the primary focus at the expense of retention and transfer of learning.

Also related to macro and micro objectives in the sequential step-by-step building of a lesson is the large number of objectives that may result. To individualize in-

struction, behavioral objectives could be developed for each student to meet his unique needs. Any significant behavioral act in learning involves a myriad of contributory acts. These are not always easy to identify in the total matrix of behavior. Furthermore, they may vary from situation to situation. To identify all behavioral acts may be impossible. Even if they could be listed, the sheer number alone, added to those behavioral objectives suited to an individual student's needs, would bring the quantity to an unworkable figure. And yet to the extent that these acts are unspecified, the behavior is left undefined and gaps may exist in the sequence hoped to contribute to the achievement of the behavioral objectives. Thus, there may be a weakening in the learning process.

Finally, the emphasis is on behavior alone. This writer does not recognize what beyond behavior is being measured. Some psychological theories suggest that behavior is important only as it sheds light on the capacities and the knowledge acquired by the student. One must draw inferences about what the student knows from his observable behavior. This seems a weakness to me in the basic assumptions of behavioral objectives. A problem exists because there is no simple and easily definable relationship between behavior and knowledge. Often behavior is only an indirect expression of knowledge. Knowledge can find expression in a variety of behaviors and the set of behaviors may be large. The fact that knowledge cannot be translated to a particular class of behaviors at the time it

is demanded by the tester or evaluator does not mean that knowledge was not acquired. It might be available at some other time or in some other form. Also, an appropriate behavioral objective is one that communicates a single behavior or single class of homogeneous behaviors associated with it. However, there are the related issues of defining units of behavior and providing criteria for determining equivalent classes of behavior.

Finally, as with the construction of a CRT, there is the problem of the expertise involved in stating clear, precise, specific, meaningful and relevant behavioral objectives. Commercial test publishers have a larger store of behavioral objectives for the various subject fields than they do criterion-referenced tests, but again the particular needs of a school district may not exactly coincide with these objectives and teachers may resist packaged objectives for their courses. Teachers, in general, are not skillful in their writing of behaviorally stated objectives.

Competency-based education in the form of criterion-referenced tests based upon behavioral objectives has been in the process of developing for a decade and the seeds appear to be taking root and some of the flowers are already blossoming. The gardens are growing and spreading. The trend, in fact, in the next several years may reshape the high schools in every state in every way: philosophy, goals, courses, teaching methods, diagnosis, evaluation, process, hours, units, attendance, and academic standards. I stand

firm in my conviction, however, that a "critical perspective"--an awareness of the strengths and weaknesses of such a approach to education must be recognized.

Conclusion

Certain generalizations can be made after a review of the related literature on criterion-referenced testing:

1. CRT scores assume a good match between test items and behavioral objectives that the CRT attempts to measure (content validity). To the extent that the test items do not accurately measure the objectives, the results will be inaccurate. To date, methodologies for item validation do not exist for CRT.

2. The theory on determing mastery levels is rather fragmentary and inconclusive. More developmental research studies on the topic need to be formulated and current research studies need to be organized and integrated into one comprehensive model.

3. Reliability and Validity studies on CRT are much needed.

4. CRT measurement should be based on threshold loss function rather than a squared-error loss function as has been done in classical test theory (Millman, 1974).

5. Bayesian statistical procedures should be used to improve proportion-correct estimates of mastery (Hambleton and Novick, 1973).

6. Customized criterion-referenced tests need to be

developed by local school districts by contracting with testing organizations (Popham, 1977).

With these considerations in mind for improving the status of criterion-referenced testing, this writer closes with the plea for a "critical perspective." Criterion-referenced testing is not the remedy of our educational ills. In fact, there are several significant limitations to the use of a Criterion-referenced testing program, especially as built into an educational program moving in the direction of competency-based education. My plea, simply stated, is to be aware of both the strengths and weaknesses of criterion-referenced testing.

Bibliography

Airasian, P.W. The role of evaluation in mastery learning. In J.H. Block (Ed.) Mastery learning: theory and practice. New York: Holt, Rinehart and Winston, 1971.

Aubertine, Secondary curriculum design and competency-based education. Educational Technology, 1972, 12, 38-39.

Bloom, B.J., Hastings, J.T., Madcus, G.F. Handbook on formative AND summative evaluation of student learning. New York: New York: McGraw Hill, 1971.

Block, J.H. Criterion-referenced measurements: potential. School Review, 1971 79, 289-97.

-----. (Ed.) Mastery learning: theory and practice. New York: Holt, Rinehart and Winston, 1971.

Burns, R.W. Achievement testing in competency-based education. Educational Technology, 1972, 12, 39-42.

Carroll, J. Problems of measurement related to the concept of learning for mastery. Educational Horizons, 1970, 48, No. 3, 71-80.

"Criterion-Referenced Testing," NEA Research Bulletin, 1971, 49, 35-36.

Davis, F.B. Criterion-referenced tests: a critique. Paper presented at the meeting of the American Educational Research Association, N.Y., February 1971. (ERIC Document Reproduction Service No. Ed. 050-154, 11 pp.)

Davis, F.B. & Diamond, J.J. The preparation of criterion-referenced tests. In C.W. Harris, M.C. Alkin, W.J. Popham (Eds.), Problems in criterion-referenced measurement. Los Angeles: UCLA Graduate School of Education, Center for the Study of Education, 1974.

Dziuban, C.D. and Vickery, K.V. Criterion-referenced measurement: some recent developments. Educational Leadership, 1973, 30, 438-486.

Ebel, R.L. Criterion-referenced measurement: limitations. School Review, 1971, 79, 282-288.

----. Essentials of educational measurement. Englewood Cliffs, N.J.: Prentice Hall, 1972.

Emrick, J.A. An evaluation model for mastery testing. Journal of Educational Measurement, 1971, 8, 321-326.

Glaser, G.R. Instructional technology and the measurement of learning outcomes. American Psychologist, 1963, 18, 519-21.

------, Nitko, A.J. Measurement in learning and instruction. In R.L. Thorndike (Ed.) Educational Measurement. Washington: American Council of Education, 1971.

Gronlund, N.E. Individualizing classroom instruction. New York: Macmillan Publishing Co., 1974.

Hambleton, R.K. Testing and decision-making procedures for selected individualized programs. Review of Educational Research, 1974, 44, 371-400.

Hambleton, R.K. and Novick, M.R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C.W. An interpretation of Livingston's coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9, 27-29.

Livingston, S.A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9,13-26.

McClendon, P.E. Criterion-referenced marking system. Educational Technology, 1976, 16, 23-25.

Meskauskas, J.A. Evaluation Models for criterion-referenced testing: view regarding mastery and standard-setting. Review of Educational Research, 1976, 46, 133-158.

Millman, J. Criterion-referenced measurement. Ch. 6 in W.J. Popham (ED.) Evaluation in education: current practices. Berkeley, Calif.: McCutchin Publishers, 1974.

--------. Passing test scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

--------. Reporting student progress: a case for a criterion-referenced marking system. Phi Delta Kappan, 1970, 52, 226-230.

Packard, R. Models of individualized instruction: the search for a measure. Educational Technology, 1972, 12, 11-14.

Popham, W. Customized criterion-refereneced tests. Educational Leadership, 1977, 34, 258-259.

Popham, James W. and Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement 1969, 6, 1-9.

Power, M.E. The grading syndrome. Journal of Reading, 1976, 19, 568-572.

Rupley, W.H. Criterion-referenced testing. ERIC/RCS in Reading Teacher, 1975, 28, 426-427+.

Washburne, C.W. Educational measurement as a key to individualizing instruction and promotion. Journal of Educational Research, 1922, 5, 195-206.

Wittrock, M.C. and D.E. Wiley (Eds.). Evaluation of instruction. New York: Holt, Rinehart and Winston, 1970.